

Translation and Tag-based Search

Shibamouli Lahiri

(shibamouli@cse.psu.edu)

Outline

- **Brief recap**
- Translation issues
- How can we use tags?

A Short Recap

- Search Engine on Ancient Greek corpus
- Query in Ancient Greek, Modern Greek or English

A Short Recap

The screenshot shows a Windows Internet Explorer browser window with the title bar "cat - Google Translate - Windows Internet Explorer provided by Comcast". The address bar displays the URL "http://translate.google.com/translate_s?hl=en&q=cat&sl=en&tl=el". The browser's menu bar includes "File", "Edit", "View", "Favorites", "Tools", and "Help". The toolbar contains various icons for search, bookmarks, and other functions. The main content area displays the Google Translate interface. At the top, it says "Google translate" with links for "Home", "Text and Web", "Translated Search", and "Tools". Below this, the "Translated Search" section shows the search term "cat" translated to "γάτα" (gata) in Greek. It also shows the source language as "English" and the target language as "Greek". A "Translate and Search" button is present. The results section, titled "Translated results from Greek web pages", shows "Results 1 - 10 of about 264,000 for γάτα". The results are presented in two columns. The left column shows the English translation, and the right column shows the original Greek text. The first result on the left is "Cat - Wikipedia", which states that the cat (Felis catus) belongs to the family Ailouridon and is the most widespread pet. The first result on the right is "Γάτα - Βικιπαίδεια", which states that the cat (Felis catus) belongs to the family Ailouridon and is the most widespread pet. Other results include "MyCat - Let's talk about cats" and "MyCat - Ας μιλήσουμε για γάτες". The bottom of the browser window shows the Windows taskbar with various icons and the system clock displaying "11:13 PM".

cat - Google Translate - Windows Internet Explorer provided by Comcast

http://translate.google.com/translate_s?hl=en&q=cat&sl=en&tl=el

File Edit View Favorites Tools Help

Google Search Search the web Search My Homepage Upload

Google translate Home Text and Web Translated Search Tools

Translated Search

Search for: cat Translated to: γάτα - [Not quite right? Edit](#)

My language: English Search pages written in: Greek

Translate and Search

Translated results from Greek web pages Results 1 - 10 of about 264,000 for γάτα

English translation

[Cat - Wikipedia](#)
The cat (Felis catus) belongs to the family Ailouridon and is the most widespread pet. He lives in the human environment for at least...
[el.wikipedia.org/wiki/Γάτα](#) - 220k - [Cached](#)

[MyCat - Let's talk about cats](#)
The comprehensive site for Greek gatofilous and cats!
[www.mycat.gr/](#) - 81k - [Cached](#)

5 +1 reasons to NOT get a cat - MyCat Community

Original Greek- [Hide Greek results](#)

[Γάτα - Βικιπαίδεια](#)
Η γάτα (Felis catus) ανήκει στην οικογένεια των Αιλουριδών και είναι το πιο διαδεδομένο κατοικίδιο ζώο. Ζει στο περιβάλλον του ανθρώπου εδώ και τουλάχιστον ...
[el.wikipedia.org/wiki/Γάτα](#) - 220k - [Προσωρινά αποθηκευμένη](#)

[MyCat - Ας μιλήσουμε για γάτες](#)
Το πληρέστερο ελληνικό site για γάτες και γατόφιλους!
[www.mycat.gr/](#) - 81k - [Προσωρινά αποθηκευμένη](#)

5+1 λόγοι για να ΜΗΝ πάρετε γάτα - MyCat Community

Internet | Protected Mode: Off 100%

gadgets 11:13 PM

Issues Identified

- Translation/transliteration
 - Disambiguation^[3,4,5,8]
 - Evaluation^[9,10,11]
- Unit of Retrieval (pages or paragraphs)
- Query disambiguation
- Ranking
- Retrieval evaluation
- Named Entity Extraction^[6]
- Summary in three languages

Outline

- Brief recap
- **Translation issues**
- How can we use tags?

Why Google Translate does not help?

- Herodotus, "The Histories", Book 1, Chapter 1, Section 0
- Ancient Greek: "Ἡροδότου Ἀλικαρνησέως ἱστορίας ἀπόδεξις ἦδε, ὥς μήτε τὰ γενόμενα ἐξ ἀνθρώπων τῷ χρόνῳ ἐξίτηλα γένηται, μήτε ἔργα μεγάλα τε καὶ θωμαστά, τὰ μὲν Ἕλλησι τὰ δὲ βαρβάροισι ἀποδεχθέντα, ἀκλεᾶ γένηται, τὰ τε ἄλλα καὶ δι' ἣν αἰτίην ἐπολέμησαν ἀλλήλοισι."
- Human translation (A.D. Godley, Harvard University Press, 1920): "This is the display of the inquiry of Herodotus of Halicarnassus, so that things done by man not be forgotten in time, and that great and marvelous deeds, some displayed by the Hellenes, some by the barbarians, not lose their glory, including among others what was the cause of their waging war on each other."
- Google Translate: "Herodotus Alikarnisseos istoriis apodexis ide, ὥς nor τὰ contemplated ἐξ people unto χρόνῳ genitai faded, nor works great ll. THOMAS, as regards the ELLIS And the varvaroisi were accepted, aklea genitai, GFs, but also self aitiin epolemisan not they."
- Yahoo Translate (AltaVista): "H[erodotoy] A[likarnisseos] i[storiis] a[podexis] ἦ[de], ὡ[s] [mite] [t]ᾱ [genomena] ἐ[x] ἄ[nthropon] [t]ῷ [chron]ῳ ἐ[xitila] [genitai], [mite] ἔ[rga] big [te] [ka]i [thomasta], [t]ᾱ [m]è[n] Ἑ[llisi] [t]ᾱ [d]è [barbaroisi] ἄ[podechthenta], ἀ[kle]ᾱ [genitai], the [te] ἄ[lla] [ka]i [di] ἥ[n] [a]i[tiin] ἐ[polemisan] ἄ[lliloisi]."

Parallel Corpus

- Perseus Project (Ancient Greek \leftrightarrow English)
- 87 books by 30 authors (4M Greek words and 6M English words)
- Europarl^[1] has almost 30M for each language
- Koehn and Monz^[2] used 15M training words for translation ({Spanish, French, Finnish, German} \rightarrow English) and roughly 60K test words
- Literary Translation \rightarrow alignment issue^[12,13]

Open Source

- [Giza++](#)
- [Moses](#)

Other Corpora

- [Project Gutenberg](#)
- [Greek Bible](#)
- [Sacred Texts](#)

Ancient Greek Dictionary

- [Lexilogos](#)
- [Kypros](#)
- [Translatum](#)
- [Ectaco](#)

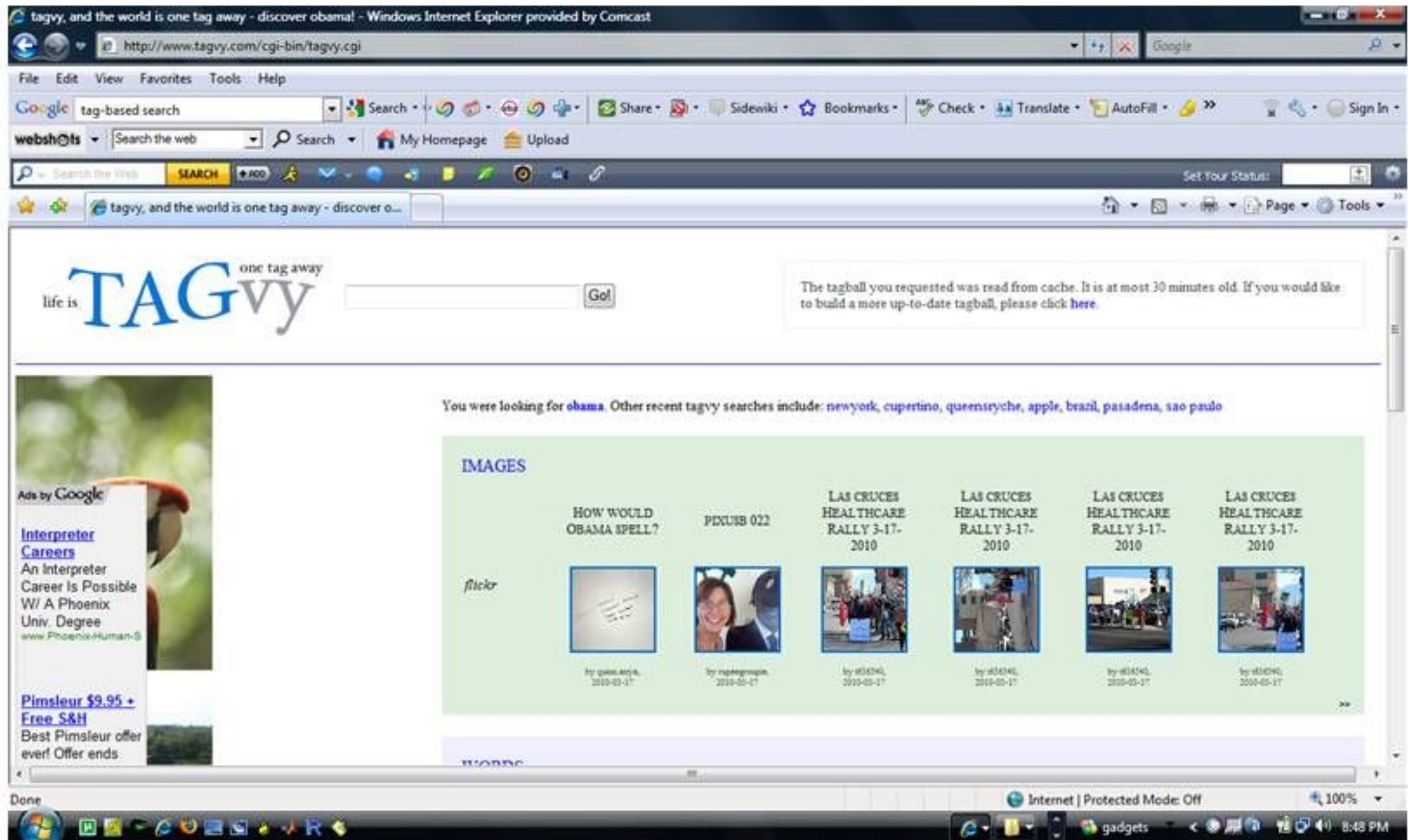
Outline

- Brief recap
- Translation issues
- **How can we use tags?**

Tag-based Search



Tag-based Search



How can tags help?

- Useful information regarding an object of interest
- Blogs, microblogs and multimedia search
- Tag clouds – the larger the tag, the higher its use
- Make “semantic” queries possible
- Query expansion^[14] and Disambiguation^[15]

Issues

- Language of tags (Greek or English)
- Level of tags (pages, paragraphs or sentences)
- Tags have their own set of problems^[7]
 - Coverage
 - Ambiguity
 - Noise
 - Spam tags

References

1. **Europarl: A Parallel Corpus for Statistical Machine Translation.** Philipp Koehn. *Machine Translation Summit (2005)*.
2. **Shared Task: Statistical Machine Translation between European Languages.** Philipp Koehn and Christof Monz. *Proceedings of the ACL Workshop on Building and Using Parallel Texts (June 2005)*.
3. **Using Structured Queries for Disambiguation in Cross-Language Information Retrieval.** David A. Hull. *AAAI (1997)*.
4. **Using the Web for Translation Disambiguation.** Y. Zhang and P. Vines. *NTCIR-5 Workshop (2005)*.
5. **Query Disambiguation for Cross-Language Information Retrieval Using Web Directories.** F. Kimura, A. Maeda, J. Miyazaki and S. Uemura. *Web Information Retrieval and Integration (2005)*.
6. **Proper name translation in cross-language information retrieval.** H.H. Chen, S.J. Huang, Y.W. Ding and S.C. Tsai. *ACL (1998)*.
7. **Information Seeking with Social Signals: Anatomy of a Social Tag-based Exploratory Search Browser.** Ed H. Chi and Rowan Nairn. *International Conference on Intelligent User Interfaces (2010)*.

References (Contd.)

8. **Iterative Translation Disambiguation for Cross-Language Information Retrieval.** Christof Monz and Bonnie J. Dorr. *In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (2005).*
9. **BLEU: a Method for Automatic Evaluation of Machine Translation.** K. Papineni. *ACL (2002).*
10. **A paraphrase-based approach to machine translation evaluation.** G. Russo-Lassner, J. Lin and P. Resnik. *(2005)*
11. **ORANGE: a Method for Evaluating Automatic Evaluation Metrics for Machine Translation.** Chin-Yew Lin and Franz Josef Och. *ACL (2004).*
12. **Text-Translation Alignment.** M. Kay and M. Röscheisen. *Computational Linguistics (1994).*
13. **HMM-based word alignment in statistical translation.** S. Vogel, H. Ney and C. Tillmann. *ACL (1996).*
14. **Query expansion using lexical-semantic relations.** E. M. Voorhees. *SIGIR (1994).*
15. **Senseval: An exercise in evaluating word sense disambiguation programs.** A. Kilgarriff. *Proceedings of LREC (1998).*